DOCUMENT RESUME

ED 419 007 TM 028 304

AUTHOR	van der Linden, Wim J.; Luecht, Richard M.
TITLE	Observed-Score Equating as a Test Assembly Problem.
INSTITUTION	Twente Univ., Enschede (Netherlands). Faculty of Educational Science and Technology.
REPORT NO	RR-97-05
PUB DATE	1997-00-00
NOTE	27p.
AVAILABLE FROM	Faculty of Educational Science and Technology, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands.
PUB TYPE	Reports - Evaluative (142)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Equated Scores; Foreign Countries; Higher Education; *Item Response Theory; Linear Programming; Observation; *Test Construction; Test Format
IDENTIFIERS	Law School Admission Test

ABSTRACT

A set of linear conditions on the item response functions is derived that guarantees identical observed-score distributions on two test forms. The conditions can be added as constraints to a linear programming model for test assembly that assembles a new test form to have an observed-score distribution optimally equated to the distribution of the old form. For a well-designed item pool, use of the model results into observed-score pre-equating and prevents the necessity of post hoc equating by a conventional observed-score equating method. An empirical example illustrates the use of the model for an item pool from the Law School Admission Test (LSAT). (Contains 6 figures and 33 references.) (Author/SLD)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED 419 007

TM

Observed-Score Equating as a Test Assembly Problem

Research Report 97-05

Wim J. van der Linden, University of Twente
and
Richard M. Luecht, National Board of Medical Examiners

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

TM028304

BEST COPY AVAILABLE

Observed-Score Equating as a Test Assembly Problem

Wim J. van der Linden
University of Twente
and
Richard M. Luecht
National Board of Medical Examiners

Abstract

A set of linear conditions on the item response functions is derived that guarantees identical observed-score distributions on two test forms. The conditions can be added as constraints to a linear programming model for test assembly that assembles a new test form to have an observed-score distribution optimally equated to the distribution of an old form. For a well-designed item pool, use of the model results into observed-score pre-equating and prevents the necessity of *post hoc* equating by a conventional observed-score equating method. An empirical example illustrates the use of the model for an item pool from the Law School Admission Test.

Observed-Score Equating as a Test Assembly Problem

A well-known method of observed-score equating is equipercentile equating. The method assumes that estimates of the observed-score distributions on the old and new test form are available and equates the observed scores on the new form to the scores on the old form estimating the same percentile in the population of examinees. With the advent of item response theory (IRT) (Lord, 1980; Rasch, 1960; Hambleton & Swaminathan, 1985; van der Linden & Hambleton, 1997), new methods of equating have become available. These methods assume that the items in the two test forms have been calibrated on the same scale for the ability parameter in the IRT model. In one method, the response functions are used to generate the observed-score distributions on both test forms, and the equipercentile method is employed to find the transformation that equates the two distributions. Another method uses the test characteristic functions of the two tests as a system of parametric equations that equates the true scores on the two tests. If the two tests have high reliability, true-score equating is often used as an approximation to observed-score equating. An introduction to equipercentile and IRT equating is given in Kolen and Brennan (1995); IRT equating is also discussed in Lord (1980).

Under IRT, tests from the same pool are automatically scored on the same scale for the ability parameter. From a theoretical point of view, it seems therefore superfluous to equate the observed-score scale as well. Nevertheless, practical reasons for this additional equating exists. Many testing programs had already fixed their score scales before IRT was introduced and replacing them by ability estimates with a more complicated relation to the response vectors than number-right scores might have been difficult to explain to their examinees. In addition, since the ability scale has a nonlinear relation to the observed-score scale, the sudden change of score distributions could have confused these examinees too. It is therefore not uncommon to find testing programs using IRT for such routines as item parameter estimation, screening of item quality, test assembly, and test equating but reporting their scores still on a traditional scale.

In these programs, tests are usually assembled from a large pool of items regularly replenished by new items calibrated on the ability scale of the pool. It is the purpose of this paper to show that, provided their item pools are well designed, such programs can omit test equating if the new test form is *assembled* to have an observed-score distribution identical to the one on the standard test to which new forms have been equated so far. As shown in this paper, the result can be obtained imposing a simple set of linear constraints on the response functions of the new test form. Use of these constraints to pre-equate observed-score distributions has several practical advantages:

1. The results hold for any population of students for which the calibration of the item pool is valid no matter its ability distribution;
2. No resources are lost running separate equating studies;
3. Scores on the new test form can be reported immediately after the test administration;
4. Unlike current equating practice, the scale of the observed scores on the new test form is not distorted by a (nonlinear) score transformation, and the scores thus keep their interpretation as number-right scores;
5. The scores on the two forms are equitable in that the procedure ensures identity of the conditional distributions of observed scores on the two forms for each possible ability level--a condition generally not met when using one of the existing equating methods (Lord, 1980, sect. 13.2).

In the remainder of the paper, first the theory of equipercentile and IRT equating is briefly reviewed. Then, a set of simple conditions on the item response function guaranteeing two test forms to have identical observed-score distributions is derived. The set replaces an earlier approximate condition given in van der Linden and Luecht (1996). The conditions are linear in the items and can be included in a linear-programming (LP) model for test assembly that optimizes the composition of the new test subject to the other test specifications already in use. Next, it will be indicated how the method can be generalized to deal with item pools dependent on more than one ability variable as well as other scoring systems than number correct. Use of the test assembly model is empirically illustrated for an item pool from the Law School Admission Test (LSAT).

Equating Transformations

The following notation is needed to present the equating transformations. Index j=1,...,n will be used to denote the items in the old test form, whereas i is used to denote the items in the new test form (i=1,...,n) or in the pool from which the form is assembled (i=1,...,I). Responses by examinee a to item i or j will be represented by random variables $U_{ai}=u_{ai}$ and $U_{aj}=u_{aj}$, respectively. Number-correct scores for examinee a on the two tests are defined as $X_a \equiv \Sigma_{i=1}^{n} U_{ia}$ and $Y_a \equiv \Sigma_{j=1}^{n} U_{aj}$, with true scores $\tau_{X_a} \equiv E(X_a)$ and $\tau_{Y_a} \equiv E(Y_a)$, respectively. Finally, it is assumed that X and Y have cumulative distribution functions F(x) and G(y).

In equipercentile equating, both test forms are administered to a single sample or two independent random samples from the population of examinees to estimate the transformation, e(x), that maps score X on the scale of score Y:

$$e(x) \equiv G^{-1}(F(x)). \quad (1)$$

The first step in this transformation identifies x as a percentile under the distribution of X; the second step equates x to the same percentile under the distribution of Y.

To discuss the mathematical equations involved in IRT equating, it is assumed that the responses to the items in the two test forms fit the 3-parameter logistic (3-PL) model. The model gives the probability of a correct response $U_{ai}=1$ as

$$P_i(\theta) \equiv c_i + (1-c_i)[1+\exp(-a_i(\theta_a - b_i))]^{-1}, \quad (2)$$

where $\theta_a \in (-\infty, \infty)$ is a parameter for the ability of examinee a, $b_i \in (-\infty, \infty)$ and $a_i \in [0, \infty)$ are parameters for the difficulty and discriminating power of item i, respectively, and $c_i \in [0,1]$ is the guessing parameter of the item. The 3-PL model is chosen because it was used to calibrate the LSAT item pool in the empirical example at the end of this paper. If $h(\theta)$ is the density of the ability distribution in the population of examinees, the probability functions of X and Y are given by

$$f(x) = \int_{-\infty}^{\infty} p_X(x|\theta)\, h(\theta)\, d\theta \quad (3)$$

and

$$g(y) = \int_{-\infty}^{\infty} p_Y(y|\theta)\, h(\theta)\, d\theta, \quad (4)$$

where $p_X(x|\theta)$ and $p_Y(x|\theta)$ are the probability functions of the conditional distributions of X and Y given θ, respectively. These conditional distributions are generalized binomial (Lord, 1980, sect. 4.1; Kendall & Stuart, 1977, sect. 5.10).

In IRT observed-score equating, the probability functions f(x) and g(y) are estimated from a random sample of examinees, estimates of the cumulative distribution functions F(x) and G(x) are calculated, and (1) is used to estimate the transformation from X to Y. Two alternative methods to implement IRT observed-score equating are discussed in Zeng and Kolen (1995).

In IRT true-score equating, the fact is used that the true scores of X and Y are equal to:

$$\tau_X = \tau_X(\theta) = \sum_{i=1}^{n} P_i(\theta) \quad (5)$$

$$\tau_Y = \tau_Y(\theta) = \sum_{i=1}^{n} P_i(\theta). \quad (6)$$

These two equations, known as the test characteristic functions of form X and Y, define the (parametric) relation between τ_X and τ_Y that can be used to equate the true score of X to the one of Y. The fact that the equations in (5)-(6) are simpler to apply than the procedure based on (3)-(4) motivates the use of true-score equating as an approximation to observed-score equating for large tests.

Conditions for Observed-Score Distributions to be Identical

From the probability functions in (3) and (4) it is clear that since the ability distribution is common, the observed-score distributions of X and Y are identical if the conditional distributions of X and Y given θ are. This fact is used in the proof of the following proposition.

Proposition 1. For any $h(\theta)$, the distributions of observed scores X and Y are identical if

$$\sum_{i=1}^{n} P_i^r(\theta) = \sum_{j=1}^{n} P_j^r(\theta), \quad \text{for } r=1,...,n. \quad (7)$$

Proof. The distribution of X given θ has no probability function in closed form but its probabilities can be obtained via the generating function $\Pi_{i=1}^{n}(Q_i + P_i)$. In addition, this probability generating function is known to have the following expansion in the powers of $P_1-\zeta$, $P_2-\zeta$, ..., $P_n-\zeta$:

$$\text{Prob}\{X=x\} = p_n(x) + \frac{n}{2} V_2 C_2(x) + \frac{n}{3} V_3 C_3(x) + (\frac{n}{4} V_4 - \frac{n^2}{8} V_2^2) C_4(x) + (\frac{n}{5} V_5 - \frac{5n^2}{6} V_2 V_3) C_5 x + \ldots, \; x=0,1,\ldots,n, \quad (8)$$

with

$$p_n(x) = \binom{n}{x} \zeta^x (1-\zeta)^{n-x}, \qquad (9)$$

$$C_r(x) = \sum_{v=0}^{r} (-1)^{v+1} \binom{r}{v} p_{n-r}(x-v), \quad r=2,\ldots,n, \qquad (10)$$

$$V_r = n^{-1} \sum_{i=1}^{n} (P_i - \zeta)^r, \quad r=2,\ldots,n, \qquad (11)$$

where ζ is defined as $n^{-1}\Sigma_{i=1}^{n} P_i$ (Walsh, 1953, 1963; Lord & Novick, 1968, sect. 23.10). Because (8) is an exact identity, the distributions of X and Y given θ are equated if the expressions in (9)-(11) are equal. For (9)-(10) this condition is realized if ζ is equal for both tests, that is, if (7) is true for r=1. In addition, (11) can be written as

$$V_r = n^{-1} \sum_{i=1}^{n} \sum_{v=0}^{r} (-1)^{r-v} \binom{r}{v} P_i^v \zeta^{r-v}, \quad r=2,\ldots,n. \qquad (12)$$

Substitution of r=2,...,n into (12) shows that these expressions are equal for both tests if (7) holds for r=1,...,n. These two conclusions establish the proposition.■

The following two propositions establish useful relations between (7) and (8).

Proposition 2. For each integer value of $R \leq n$, the set of equalities in (7) obtained for r=1,...,R equate the first R terms of the series in (8).

The truth of this proposition follows immediately from the substitution at the end of the proof of Proposition 1.■

Proposition 3. For $n \rightarrow \infty$, the contributions of the terms in (8) of the order r>2 vanishes.

Proof. For r=1, the condition in (7) formulates that the test characteristic functions in (5)-(6) be equal. Thus, equal test characteristic functions imply that the two distributions have identical first terms for (8), and true-score equating can be viewed as a first-order approximation to observed-score equating. Since the observed-score distribution converges to the true-score distribution if test length increases, this first-order approximation improves with

BEST COPY AVAILABLE

test length. As a result, for longer tests the contribution of the higher-order terms in (8) decreases.■

The practical implication of the proposition is that the series in (8) can be truncated after a few terms. The same can be done for the test assembly model below imposing (7) on the selection of the items only for small values of r. In the empirical example later in this paper only the first three terms were used.

An interesting consequence, however, is obtained if all n equalities in (7) are imposed. The set of conditions is then equivalent to the one of the response functions of the two tests being pairwise identical. This property as well as its proof were suggested by N. D. Verhelst (personal communication, November 1, 1996):

Proposition 4. The conditions in (7) hold simultaneously for all values of r if and only if the two tests have pairwise identical response functions.

Proof. If the two tests have pairwise identical response functions, then the conditions in (7) hold trivially. The proof of the reverse implication is based on the idea to define probability spaces over the two sets of response functions and to invoke the principle of moments (Kendall & Stuart, 1977, sect. 4.22). Thus let be $< \chi_\theta, \wp(\chi_\theta), p >$ a (finite) probability space, where χ is the set of response probabilities in test X for a fixed value of θ, $\wp(\chi_\theta)$ is the power set of χ_θ and p is the (uniform) probability function $p(i) \equiv 1/n$. In addition, a random variable $X_\theta(i) \equiv P_i(\theta)$ is defined. An analogous probability space and random variable is defined for test Y. The conditions in (7) stipulate that the first n moments of the distributions induced by the two random variables are identical. Therefore, the two distributions are identical; that is, for each value of $X_\theta(i)=P_i(\theta)$ there exist a value $Y_\theta(j)=P_j(\theta)$, and vice versa. Because the argument holds for an arbitrary value of θ, the pair of functions $P_i(\theta)$ and $P_j(\theta)$ have more than two points in common and are identical.■

Note that two tests with pairwise identical response functions have equal true scores and observed-score variances for each examinee in the population for which the IRT model holds. These tests therefore yield parallel measurements (Lord & Novick, 1968, definition 2.13.1). Proposition 4 thus shows the (stringent) conditions in IRT under which the classical definition of parallel measurements hold.

Proposition 4 also implies that a test assembly model with a larger number of the equalities in (7) imposed on the new test might be a convenient one-stage alternative to the two-stage approaches to item matching proposed by van der Linden and Timminga (1988) (see also Armstrong & Jones, 1992). However, as the focus of this paper is on the less

demanding problem of observed-score equating, the latter suggestion is not further elaborated here.

Test Assembly Model

Because the conditions in (7) are linear in the items, they can be used as objective function and/or constraints in an LP model for optimal test assembly. Such models have been proposed earlier, for example, to assemble tests to match a target information function (Swanson & Stocking, 1993; Theunissen, 1985; van der Linden & Boekkooi-Timminga, 1989), to assemble sets of parallel test forms (Adema, 1992; Armstrong, Jones & Wu, 1992 ; Boekkooi-Timminga, 1987, 1990), to maximize classical test reliability (Adema & van der Linden, 1989; Armstrong, Jones & Wang, 1994), to match tests item by item (Armstrong & Jones, 1992; van der Linden & Boekkooi-Timminga, 1988), or to implement constrained adaptive testing (van der Linden & Reese, in press). In addition, these models allow for all other test specifications typically constraining the selection of items in a testing program.

Following is the model proposed to select a new test form from a pool of items with an observed-score distribution optimally equated to the distribution of an old form. Let x_i, i=1,...,I, be the decision variables to denote whether (x_i=1) or not (x_i=0) item i is included in the new test form. Because the first terms in (8) are most important, the idea is to choose the values of x_i such that the differences between the two left-hand and right-hand sums in (7) are minimized for r=1,...,R at θ_k , k=1,...,K. As already noted R can be small. Also, since the item response functions in (7) are well-behaved continuous functions, only a few points are necessary to control their shapes over the range of θ values considered. However, there are no limitations as to the number of values and their spacing, and test assemblers are free to select the set best fitting their needs.

The model is as follows:

minimize y **(13)**

subject to

$$\sum_{i=1}^{I} P_i^r(\theta_k)\, x_i - \sum_{j=1}^{n} P_j^r(\theta_k) \le y, \qquad k=1,...,K;\ r=1,...,R \qquad \textbf{(14)}$$

$$\sum_{i=1}^{I} P_i^r(\theta_k)\, x_i - \sum_{j=1}^{n} P_j^r(\theta_k) \ge -y, \qquad k=1,...,K;\ r=1,...,R \qquad \textbf{(15)}$$

$$\sum_{i=1}^{I} x_i = n, \tag{16}$$

$$\sum_{i=1}^{I} q_{is} x_i \leq r_s, \qquad s=1,\ldots,S, \tag{17}$$

$$\sum_{i \in V_s} x_i = n_s, \qquad s=1,\ldots,S, \tag{18}$$

$$x_i \in 0,1, \qquad i=1,\ldots,I, \tag{19}$$

$$y \geq 0, \tag{20}$$

The constraints in (14)-(15) require the difference between $\Sigma_{i=1}^{I} P_i^r(\theta_k) x_i$ and $\Sigma_{j=1}^{n} P_j^r(\theta_k)$ to be in the interval [-y,y], $y \geq 0$, for k=1,...,K and r=,...,R whereas the objective function in (13) minimizes y. The model therefore effectively minimizes the largest difference between these sums over the θ values selected and is thus of the minimax type. The constraint in (16) sets the length of the new test form equal to n. The constraints in (17)-(18) deal with all possible additional test specifications. For example, if the total length of the test measured by its number of lines should be smaller than a given number r_s, q_{is} can be defined as the number of lines in item i, and the constraint in (17) guarantees the result. Likewise, if some of the items in the pool measure the application of certain skills, the set V_s in (18) can be chosen to be this subset of items and the constraint guarantees the selection of n_s items from it. Various other types of constraints are possible; for a review see van der Linden (in press) or van der Linden and Boekkooi-Timminga (1989). Finally, the constraints in (19)-(20) define the range of the decision variables.

As already noted, the series in (8) approximates the distribution generally good for only a few terms but that the precision of the results goes up if the upper bound R in (14)-(15) is increased. This result is only analytical, however; the actual problem is one of combinatorial optimization. In practice, however item pools are finite and not all possible combinations of values for the item parameters are available. As a consequence, imposing too many of the conditions in (7) may occasionally lead to item combinations compromising between the conditions with results slightly worse than those for a case of fewer conditions. Though weights could be added to the right-hand side variables in the individual constraints in (14)-(15), it is hard to base the choice of their values on a theoretic argument. In the empirical example below it proved best to include constraints for the first two or three terms in (8) in the model and to apply no weighing.

BEST COPY AVAILABLE

The above model can be solved for optimal values of x_i and y using standard LP software or the test assembly software package ConTEST (Timminga, van der Linden & Schweizer, 1996). For models with a special structure, heuristics as in Luecht and Hirsch (1992) are convenient. The choice of algorithm to solve the model is further addressed in the presentation of the empirical example below.

Multidimensional Ability

A potential danger to IRT-based equating is lack of fit of the response model to the data. Such lack of fit is most likely due to the fact that success on the item pool can be dependent on more than one ability. An obvious remedy is to use a multidimensional response model. A well-known model is the following extension of the 2-PL model, i.e., the model in (2) with $c_i=1$ for i=1,...,n:

$$P_i(\theta) \equiv P(U_i = 1|\theta_1,...,\theta_D, a_i1,...,a_iD, b_i) \equiv \frac{\exp(\sum_{d=1}^{D} a_{id}\theta_d - b_i)}{1+\exp(\sum_{d=1}^{D} a_{id}1\theta_d - b_i)}, \qquad (21)$$

where θ_d, d=1,...,D, are the ability variables, a_{id} is the parameter for the discriminating power of item i along θ_d, and b_i is a parameter for the composite difficulty of item i. Detailed information about the model is given in McKinley and Reckase (1983), Reckase (1985, 1997), and Samejima (1974). To equate tests measuring possible multidimensional abilities, Glas (1992) uses a multidimensional Rasch or 1-PL model. The model is equivalent to the one in (21) with $a_{id}=1$ for all i and d, but assumes that the items display a "simple structure" with respect to their dependencies on the ability variables; that is, the success on disjoint subsets of items in the pool is modeled as being dependent on different unidimensional ability variables. In addition, the individual abilities are linked by the assumption of a multivariate normal distribution for the population of examinees.

Test assembly from a multidimensional item pool requires a slight adaptation of the model. The only changes necessary are substituting the multidimensional response functions into the conditions in (7) and specifying a multidimensional rather than a unidimensional grid of ability values for the constraints. That is, (14)-(15) has to be replaced by

$$\sum_{i=1}^{I} P_i^r(\theta_{dk})\, x_i - \sum_{j=1}^{n} P_j^r(\theta_{dk}) \le y, \qquad d=1,...,D;\ k=1,...,K;\ r=1,...,R. \qquad (22)$$

$$\sum_{i=1}^{I} P_i^r(\theta_{dk})\, x_i \;-\; \sum_{j=1}^{n} P_j^r(\theta_{dk}) \;\geq\; -y, \qquad d=1,...,D;\ k=1,...,K;\ r=1,...,R. \qquad (23)$$

In the model by Glas, the grids remain unidimensional but different grids have to be specified for the separate ability variables for the different subsets of items in the pool.

Other Scores than Number Correct

Test results are often reported as a conversion of the number-correct score. If the conversion is a monotonically increasing function, the test assembly model in this paper can just be applied to the number-correct scores which are then converted to the desired scale afterwards. Examples of conversions to which this principle applies are changes of origin and/or unit of number-correct scores and "formula scoring" to correct for possible random guessing on multiple-choice items.

Empirical Example

The test assembly model was applied to a former pool of 753 items from the Law School Admission Test (LSAT) program. The items in the pool were calibrated using the 3-PL model in (2). The pool consisted of items falling into three different content categories, labeled SA, SB, and IA here. In addition, items varied in (sub)type, gender and minority orientation, answer key, and word count. Finally, a portion of the item pool had a set structure with items in the same set sharing a common stimulus. The type of stimulus varied in content description. All existing specifications for the LSAT were modeled as linear constraints following the general format in (17)-(18). To model the inclusion of item sets in the test, a second type of decision variables was needed in addition to the variables x_i in (13)-(20). In all, the model had 729 variables and 433 constraints. An old test assembled by hand to meet several specifications of the LSAT was known to the authors. The model in (13)-(20) was used to assemble new tests of 75 items with observed-score distribution optimally equated to the distribution on the old test.

The model was solved using the First Acceptable Integer Solution heuristic as implemented in the ConTEST program. The heuristic first calculated an upper bound to the value of the objective function in (13) relaxing the other decision variables in the model and then performs a branch-and-bound search for the optimal solution that is stopped when the first integer solution with objective function value within a small tolerance from the upper bound is found. The search is speeded up using the optimal reduced costs in the solution to the relaxed model to fix some of the decision variables. For further details, see Timminga, van der Linden and Schweizer (1996, sect. 6.6.5). In the current application, the search for a 0-1

BEST COPY AVAILABLE

solution was stopped as soon as the value of the objective function, y, was smaller than .01. The observed-score distributions for the old and new tests were generated according to (3)-(4), with θ distributed as N(0,1), using a recursive algorithm introduced in Lord and Wingersky (1984).

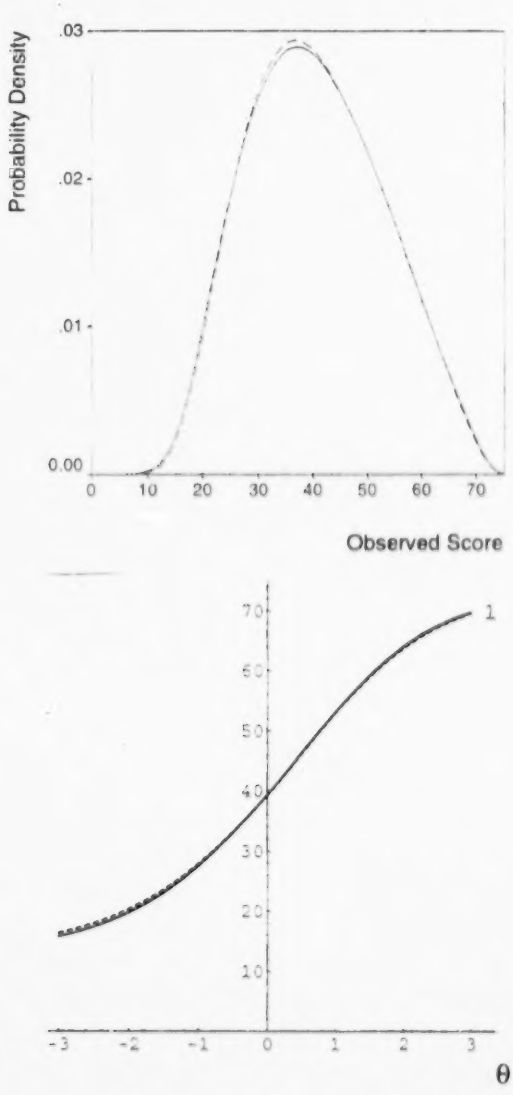


Figure 1. Probability functions of observed-score distributions (upper panel) and condition in (7) (lower panel) for r=1 and θ=0 (solid line: new test; dashed line: old test).

BEST COPY AVAILABLE

Two sets of results were calculated, both for r=1,...,3 but one with the response functions controlled only at $\theta=0$ and the other at $\theta_1=-1.0$ and $\theta_2=1.0$. The probability functions of the observed-score distributions for the first set are plotted in Figures 1-3.

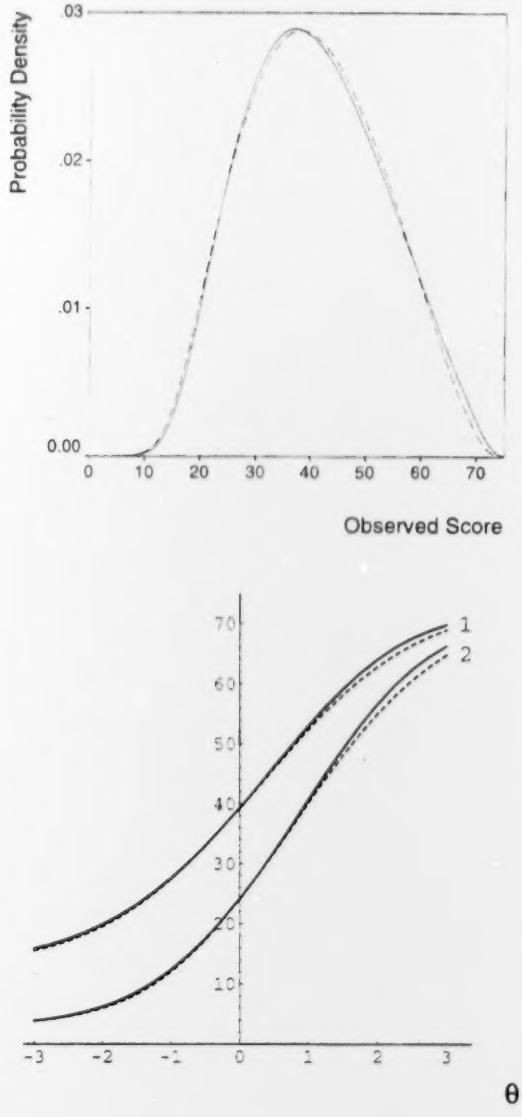


Figure 2. Probability functions of observed-score distributions (upper panel) and conditions in (7) (lower panel) for r=1,2 and $\theta=0$ (solid line: new test; dashed line: old test).

BEST COPY AVAILABLE

The figures also show the extent to which the the sums of the rth power of the response functions for the old and new tests are equal.

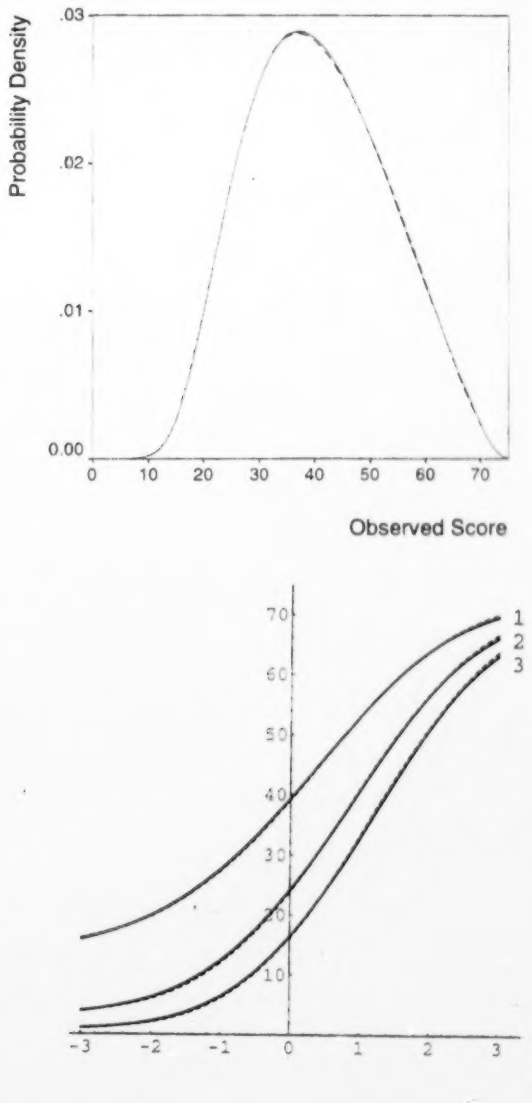


Figure 3. Probability functions of observed-score distributions (upper panel) and conditions in (7) (lower panel) for r=1,2,3 and θ=0 (solid line: new test; dashed line: old test).

BEST COPY AVAILABLE

The general impression is that the probability functions of the old and new tests are nearly identical in all four cases, with perfect results for r=1,2,3. In this case, the sums of the three powers of the response functions are also identical for all practical purposes. The sets of curves for second case are given in Figures 4-6.

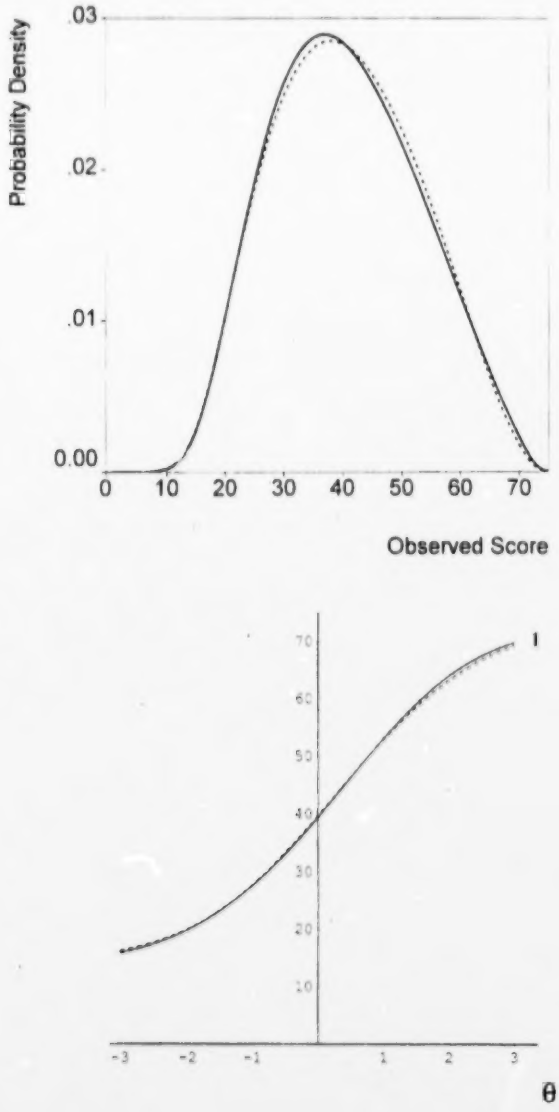


Figure 4. Probability functions of observed-score distributions (upper panel) and condition in (7) (lower panel) for r=1 and θ_1=-1.0 and θ_2=1.0 (solid line: new test; dashed line: old test).

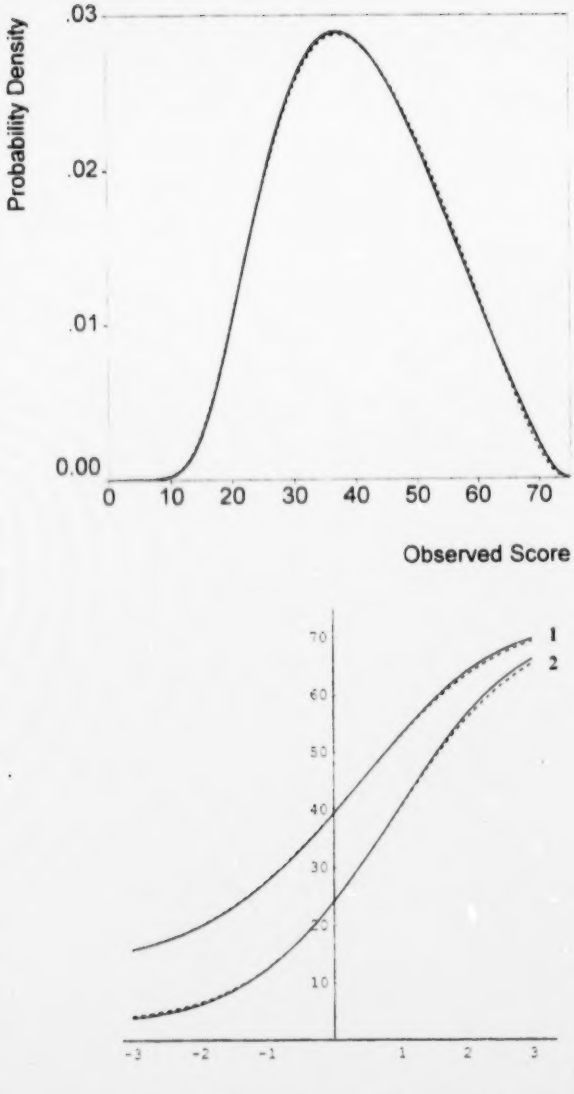
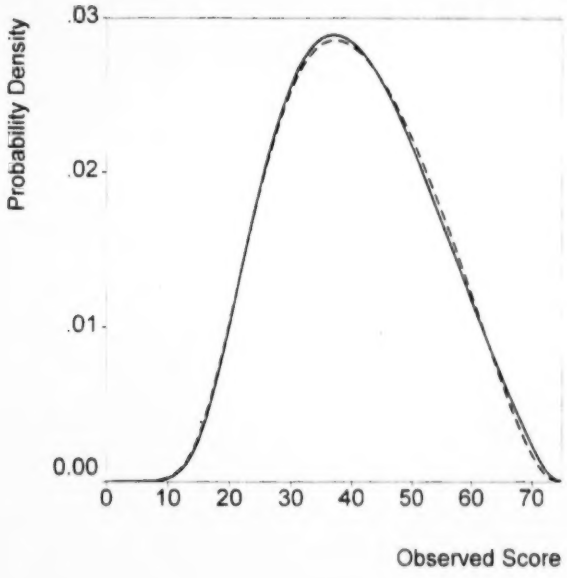


Figure 5. Probability functions of observed-score distributions (upper panel) and conditions in (7) (lower panel) for r=1,2 and θ_1=-1.0 and θ_2=1.0 (solid line: new test; dashed line: old test).



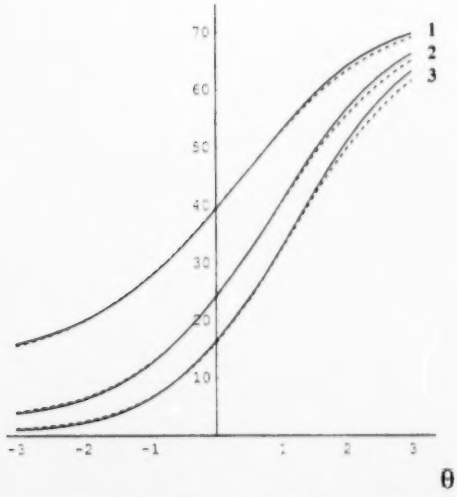


Figure 6. Probability functions of observed-score distributions (upper panel) and conditions in (7) (lower panel) for r=1,2,3 and θ_1=-1.0 and θ_2=1.0 (solid line: new test; dashed line: old test).

For r=1 the fit is comparable to the one obtained for the previous case. The only change is a shift of the new distribution lightly to the left. For r=1,2 the results are perfect. The slight

BEST COPY AVAILABLE

decrease in fit for r=1,2,3 is due to the size of the item pool. As noted earlier, if the item pool does not have all possible combinations of parameter values, the goal to find a good compromise between all conditions may lead to worse result, in particular if the new condition, such as the one for r=3, only has a slight impact on the shape of the observed-score distribution.

Discussion

The success of the test assembly model proposed in this paper is predicated on the quality of the item pool. If the pool is small relative to the size of the test or not well designed, the observed score distribution of the test assembled from the pool may fit the distribution of target test not as well as in the empirical example above. However, in such cases the use of the test assembly model is still recommended; its results guarantee that the additional transformation necessary to equate the two test forms exactly involves a minimal distortion of scale over all possible test forms from the pool. Whether or not additional equating is necessary can immediately be inferred from such output as in Figures 1-6. The observed-score distributions of the two test forms in these plots are all that is needed to perform additional equipercentile equating. This equating can take place before the test is administered.

The quality of the item pool is also determined by the quality of item calibration and the fit of the (unidimensional or multidimensional) IRT model. The robustness of the results in this paper against item calibration errors or item misfit has not been examined yet.

References

Adema, J.J. (1990). The construction of customized two-staged tests. *Journal of Educational Measurement*, *27*, 241-253.

Adema, J.J. (1992). Methods and models for the construction of weakly parallel tests. *Applied Psychological Measurement*, *16*, 53-63.

Adema, J.J. & van der Linden, W.J. (1989). Algorithms for computerized test construction using classical item parameters. *Journal of Educational Statistics*, *14*, 279-290.

Armstrong, R.D. and Jones, D.H. (1992). Polynomial algorithms for item matching. *Applied Psychological Measurement*, *16*, 365-373.

Armstrong, R.D., Jones, D.H., & Wang, Z. (1994) Automated parallel test construction using classical test theory. *Journal of Educational Statistics*, *19*, 73-90.

Armstrong, R.D., Jones, D.H., & Wu, I-L. (1992). An automated test development of parallel tests. *Psychometrika*, *57*, 271-288.

Boekkooi-Timminga, E. (1987). Simultaneous test construction by zero-one programming. *Methodika*, *1*, 1101-112.

Boekkooi-Timminga, E. (1990). The construction of parallel tests from IRT-based item banks. *Journal of Educational Statistics*, 15, 129-145.

Glas, C.A.W. (1992). A Rasch model with a multivariate distribution of ability. In M. Wilson (Ed.), *Objective measurement: Theory into practice* (Vol. 1). Norwood, NJ: Ablex.

Hambleton, R.K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston: Kluwer-Nijhoff.

Kendall, M.G., & Stuart, A. (1977). *The advanced theory of statistics* (Vol. 1, 4th ed.). London: Griffin & Co.

Kolen, M.J., & Brennan, R.L. (1995). *Test equating: Methods and practices*. New York: Springer-Verlag.

Lord, F.M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Erlbaum.

Lord, F.M., & Wingersky, M.S. (1984). Comparison of IRT true-score and equipercentile observed-score "equatings". *Applied Psychological Measurement*, *8*, 452-461.

Luecht, R.M. & Hirsch, T.M. (1992). Computerized test construction using average growth approximation of target information functions. *Applied Psychological Measurement*, *16*, 41-52.

McKinley, R.L., & Reckase, M.N. (1983). *An extension of the two-parameter logistic model to the multidimensional latent space* (Research Report ONR 83-2). Iowa City, IA:American College Testing.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests.* Copenhagen: Danmarks Paedagogiske Institut.

Reckase, M.D. (1985). The difficulty of test items that measure more than one ability. *Applied Psychological Measurement, 9*, 401-412.

Reckase, M.D. (1997). A linear logistic multidimensional model for dichotomous item response data. In W.J. van der Linden & R.K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 271-286). New York City, NY: Springer-Verlag.

Samejima, F. (1974). Normal ogive model for the continuous response level in the multidimensional latent space. *Psychometrika, 39*, 111-121.

Swanson, L. & Stocking, M.L. (1993). A model and heuristic for solving very large item selection problems. *Applied Psychological Measurement, 17*, 151-166.

Theunissen, T.J.J.M. (1985). Binary programming and test design. *Psychometrika, 50*, 411-420.

Timminga, E. & van der Linden, W.J., & Schweizer, D.A. (1996). *ConTEST 2.0: A decision support system for item banking and optimal test assembly* (computer program and manual). Groningen, The Netherlands: iec ProGAMMA.

van der Linden, W.J. (1996). Assembling test for the measurement of multiple traits. *Applied Psychological Measurement, 20*, 373-388.

van der Linden, W.J. (in press). Optimal assembly of educational and psychological tests, with a bibliography. *Applied Psychological Measurement*.

van der Linden, W.J., & Boekkooi-Timminga, E. (1988). A zero-one programming approach to Gulliksen's matched random subsets method. *Applied Psychological Measurement, 12*, 201-209.

van der Linden, W.J., & Boekkooi-Timminga, E. (1989). A maximin model for test design with practical constraints. *Psychometrika, 17*, 237-247.

van der Linden, W.J. & Hambleton, R.K. (Eds.) (1997). *Handbook of modern item response theory*. New York City, NY: Springer-Verlag.

van der Linden, W.J. & Luecht, R.M. (1966). An optimization model for test assembly to match observed-score distributions. In G. Engelhard & M. Wilson (Ed.), *Objective measurement: Theory into practice* (Vol.3, pp. 405-418). Norwood, New Jersey: Ablex Publishing Company.

van der Linden, W.J., & Reese, L.M. (in press). A model for optimal constrained adaptive testing. *Applied Psychological Measurement*,

Walsh, J.E. (1953). Approximate probability values for observed number of successes. *Sankhya, 15*, 281-290.

Walsh, J.E. (1963). Corrections to two papers concerned with binomial events. *Sankhya*, Series A, *25*, 427.

BEST COPY AVAILABLE

Zeng, L., & Kolen, M.J. (1995). An alternative approach for IRT observed-score equating of number-correct scores. *Applied Psychological Measurement, 19*. 231-241.

Authors' Note

The authors are most indebted to Norman D. Verhelst for suggesting Proposition and its proof, to the Law School Admission Council (LSAC) for making available the data set, and to Wim M.M. Tielen for his computational assistance. Correspondence should be send to W.J. van der Linden, Department of Educational Measurement and Data Analysis, University of Twente, P.O. Box 217, 7500 AE Enschede, The Netherlands. Email: vanderlinden@edte.utwente.nl

Titles of Recent Research Reports from the Department of Educational Measurement and Data Analysis. University of Twente, Enschede, The Netherlands.

RR-97-05 W.J. van der Linden & Richard M. Luecht, *Observed-Score Equating as a Test Assembly Problem*

RR-97-04 W.J. van der Linden & J.J. Adema, *Simultaneous Assembly of Multiple Test Forms*

RR-97-03 W.J. van der Linden, *Multidimensional Adaptive Yesting with a Minimum Error-Variance Criterion*

RR-97-02 W.J. van der Linden, *A Procedure for Empirical Initialization of Adaptive Testing Algorithms*

RR-97-01 W.J. van der Linden & Lynda M. Reese, *A Model for Optimal Constrained Adaptive Testing*

RR-96-04 C.A.W. Glas & A.A. Béguin, *Appropriateness of IRT Observed Score Equating*

RR-96-03 C.A.W. Glas, *Testing the Generalized Partial Credit Model*

RR-96-02 C.A.W. Glas, *Detection of Differential Item Functioning using Lagrange Multiplier Tests*

RR-96-01 W.J. van der Linden, *Bayesian Item Selection Criteria for Adaptive Testing*

RR-95-03 W.J. van der Linden, *Assembling Tests for the Measurement of Multiple Abilities*

RR-95-02 W.J. van der Linden, *Stochastic Order in Dichotomous Item Response Models for Fixed Tests, Adaptive Tests, or Multiple Abilities*

RR-95-01 W.J. van der Linden, *Some decision theory for course placement*

RR-94-17 H.J. Vos, *A compensatory model for simultaneously setting cutting scores for selection-placement-mastery decisions*

RR-94-16 H.J. Vos, *Applications of Bayesian decision theory to intelligent tutoring systems*

RR-94-15 H.J. Vos, *An intelligent tutoring system for classifying students into Instructional treatments with mastery scores*

RR-94-13 W.J.J. Veerkamp & M.P.F. Berger, *A simple and fast item selection procedure for adaptive testing*

RR-94-12 R.R. Meijer, *Nonparametric and group-based person-fit statistics: A validity study and an empirical example*

RR-94-10 W.J. van der Linden & M.A. Zwarts, *Robustness of judgments in evaluation research*

RR-94-9 L.M.W. Akkermans, *Monte Carlo estimation of the conditional Rasch model*

RR-94-8 R.R. Meijer & K. Sijtsma, *Detection of aberrant item score patterns: A review of recent developments*

RR-94-7 W.J. van der Linden & R.M. Luecht, *An optimization model for test assembly to match observed-score distributions*

RR-94-6 W.J.J. Veerkamp & M.P.F. Berger, *Some new item selection criteria for adaptive testing*

RR-94-5 R.R. Meijer, K. Sijtsma & I.W. Molenaar, *Reliability estimation for single dichotomous items*

RR-94-4 M.P.F. Berger & W.J.J. Veerkamp, *A review of selection methods for optimal design*

RR-94-3 W.J. van der Linden, *A conceptual analysis of standard setting in large-scale assessments*

RR-94-2 W.J. van der Linden & H.J. Vos, *A compensatory approach to optimal selection with mastery scores*

RR-94-1 R.R. Meijer, *The influence of the presence of deviant item score patterns on the power of a person-fit statistic*

RR-93-1 P. Westers & H. Kelderman, *Generalizations of the Solution-Error Response-Error Model*

RR-91-1 H. Kelderman, *Computing Maximum Likelihood Estimates of Loglinear Models from Marginal Sums with Special Attention to Loglinear Item Response Theory*

RR-90-8 M.P.F. Berger & D.L. Knol, *On the Assessment of Dimensionality in Multidimensional Item Response Theory Models*

RR-90-7 E. Boekkooi-Timminga, *A Method for Designing IRT-based Item Banks*

RR-90-6 J.J. Adema, *The Construction of Weakly Parallel Tests by Mathematical Programming*

RR-90-5 J.J. Adema, *A Revised Simplex Method for Test Construction Problems*

RR-90-4 J.J. Adema, *Methods and Models for the Construction of Weakly Parallel Tests*

RR-90-2 H. Tobi, *Item Response Theory at subject- and group-level*

RR-90-1 P. Westers & H. Kelderman, *Differential item functioning in multiple choice items*

Research Reports can be obtained at costs, Faculty of Educational Science and Technology, University of Twente, Mr. J.M.J. Nelissen, P.O. Box 217, 7500 AE Enschede, The Netherlands.